The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion*

Marc F. Bellemare† Jeffrey R. Bloem‡ Noah Wexler§

September 11, 2020

Abstract

We present the first application of Pearl's (1995) front-door criterion to observational data wherein the required point-identification assumptions plausibly hold. For identification, the front-door criterion exploits exogenous mediator variables on the causal path. We estimate the effect of authorizing a shared Uber or Lyft ride on tipping by exploiting the plausibly exogenous variation in whether one actually shares a ride with a stranger conditional on authorizing sharing, on fare level, and on time-and-place fixed effects. We find that most of the observed negative effect on tipping is driven by selection. We then explore the consequences of violating the identification assumptions.

Keywords: Front-Door Criterion, Causal Inference, Causal Identification, Treatment Effects, Ride-Hailing

JEL Codes: C13, C18, R40, D90

*We thank Chris Auld, David Childers, Carlos Cinelli, Dave Giles, Paul Glewwe, Adam Glynn, Paul Hünermund, Guido Imbens, Jason Kerwin, Daniel Millimet, Judea Pearl, Bruce Wydick, and seminar participants at the World Bank and Michigan State University for useful comments and suggestions. This research was conducted prior to Jeffrey R. Bloem's employment at the USDA. The findings and conclusions in this manuscript are those of the authors and should not be construed to represent any official USDA or US Government determination or policy. All remaining errors are ours.

†Corresponding Author. Northrop Professor, Department of Applied Economics, University of Minnesota, 1994 Buford Avenue, Saint Paul, MN 55108, Email: mbellema@umn.edu.

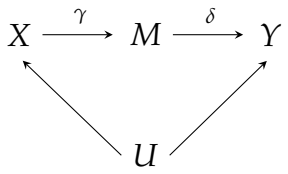
‡Research Economist, United States Department of Agriculture, Economic Research Service, MS9999, Beacon Facility, P. O. Box 419205, Kansas City, MO 64141, Email: Jeffrey.Bloem@usda.gov.

§Ph.D. Student, Humphrey School of Public Affairs, University of Minnesota. 301 19th Ave. S, Minneapolis, MN 55455, Email: wexle059@umn.edu.

1 Introduction

We present the first application of Judea Pearl's (1995, 2000) front-door criterion (FDC) to observational data in which the required assumptions for point-identification plausibly hold. The directed acyclic graph (DAG) in Figure I illustrates the FDC setup,[1] where the reduced-form relationship between outcome variable Y and treatment variable X is biased because of the presence of unobserved confounders U, which cause both X and Y.

FIGURE I: The Front-Door Criterion



Pearl's insight is that when there exists a mediator variable M on the causal path from X to Y and that mediator is not directly caused by U, it is possible to estimate the average treatment effect (ATE) of X on Y.[2] This is done by (i) estimating the effect γ of X on M (which is identified because the unobserved confounders in U cause X but not M), (ii) estimating the effect δ of M on Y conditional on X (which is identified because the unobserved confounders in U cause Y but not M), and (iii) multiplying the estimates $\hat{\gamma}$ and $\hat{\delta}$ by each other. This last step yields the ATE of X on Y, which we label $\hat{\beta}$ to keep with convention. Intuitively, the FDC estimates the ATE because it decomposes a reduced-form relationship that is not causally identified into two causally identified

[1] For readers unfamiliar with DAGs, directed arrows (i.e., $\rightarrow$) represent causal relationships between variables. In a DAG, $X \rightarrow Y$ simply means that $Y = f(X, e_Y)$, where e_Y is an error variable independent of either X or Y. In a DAG, the causal relationship $X \rightarrow Y$ flowing from X to Y need not be parametric or even linear (Morgan and Winship 2015), but we focus in this paper on parametric, linear relationships for simplicity.

[2] Although the literature often refers to M as a mechanism, we will refer to it as mediator in this paper. Our view is that the "mechanism" terminology is more accurate when discussing the theoretical framework underlying an application of the front-door criterion, and the "mediator" terminology is more accurate when discussing the statistical setup of the front-door criterion, as we do in this paper. Pearl (1993) originally called M a "mediating instrumental variable" and described M as an "exogenously-disturbed mediator."

relationships.[3]

Despite its relative simplicity, economists have been reluctant to incorporate the front-door criterion in their empirical toolkit. Anecdotally, that resistance appears to stem from the fact that finding a convincing empirical application has thus far proven elusive (see, e.g., Imbens 2020; Gupta et al. 2020). We provide such an application and further address the following questions: How can the front-door criterion be used in the context of linear regression? And what happens when the necessary identification assumptions for the front-door criterion to estimate an ATE do not hold strictly?

In his writings on the front-door criterion, Pearl repeatedly provides the same example of an empirical application. In his canonical example, X is a dummy variable for whether one smokes, Y is a dummy variable for whether one develops lung cancer, and M is the accumulation of tar in one's lungs (Pearl 1995; Pearl 2000; Pearl and Mackenzie 2018). But some have pointed out that if (i) smoking has a direct effect on lung cancer, independent on tar accumulation or (ii) both tar accumulation and lung cancer are caused by alternative sources, such as a hazardous work environment, then this canonical example violates the necessary FDC identifying assumptions (see, e.g., Imbens 2020). Consequently, the adoption of the FDC has been slow among applied researchers. The only extant published social science applications of the FDC are by Glynn and Kashin (2017; 2018).[4] We build on these previous contributions, although it is important to note that the authors of those previous studies themselves admit that the necessary assumptions required for credible identification with the FDC approach do not hold.[5]

[3]Some readers may be tempted to use the mediator M as an instrument for the endogenous treatment X. In Appendix 1, we explain why this is not desirable.

[4]In Glynn and Kashin (2018) the authors apply the FDC approach to estimate the effects of attending a job training program on earnings. In Glynn and Kashin (2017) the authors also apply the FDC difference-in-differences approach to evaluate the effect of an early in-person voting program on voter turnout. Both applications closely approximate, but ultimately do not exactly replicate existing experimental estimates.

[5]Specifically, Glynn and Kashin (2018) write, "As we discuss in detail below, the assumptions implicit in [the FDC] graph will not hold for job training programs, but this presentation clarifies the inferential approach." In Glynn and Kashin (2017), the authors develop a difference-in-difference extension to the FDC approach which requires an exclusion restriction and a parallel trends assumption specifically for their empirical setting where the necessary conditions for the FDC do not hold. This previous work was helpful in partially identifying the treatment effect of interest, thereby establishing reasonable bounds on

Our contribution is threefold. First, because linear regression remains the workhorse of applied economics and an explanation of how to use the front-door criterion in a regression context has so far been lacking in the literature, we explain how to estimate treatment effects with the front-door criterion in the context of linear regression. Estimation relies on Pearl's three identification assumptions and an additional empirically verifiable requirement of one-sided noncompliance.[6]

Second, we present two examples of the FDC in practice. One uses simulated data to show an ideal application of the FDC—one in which we know the true ATE.[7] Our second example is the core contribution of this paper because it presents the first application of the FDC to observational data where the necessary assumptions for point-identification of the ATE plausibly hold. In that application, we estimate the effect on tipping behavior of authorizing a ride-hailing app such as Lyft or Uber to overlap a ride with another paying passenger. We find that the observed negative correlation between choosing to share a ride and tipping is almost entirely explained by selection into treatment—a finding relevant to the economics of tipping (Azar 2020).

In our application, we are not able to randomly assign whether someone authorizes shared rides (i.e., X) on Uber or Lyft. Additionally, since the base fare of shared rides is typically less than that of solo rides, this choice is clearly endogenous to tipping behavior (i.e., Y). Once a passenger chooses to share a ride, however, they will not necessarily share a ride. We can therefore exploit the exogenous variation—conditional on fare level and date, hour, day of the week–hour, and origin–destination fixed effects—in whether or not a passenger actually shares a ride (i.e., M) once they authorize sharing. In that case, the front-door criterion can credibly estimate the causal effect of authorizing shared rides on tipping behavior (i.e., Y).

effect estimates. In contrast, the treatment effects we estimate here are point-identified.

[6]This contribution builds on the previous work by Chalak and White (2011).

[7]This simulated example will serve as the basis for our third contribution below, where we explore departures from the necessary assumption for the front-door criterion to yield the average treatment effect of X on Y.

Third, and perhaps most importantly for applied researchers, we explore what happens when the necessary assumptions for the front-door criterion to identify the ATE of X on Y fail to hold. Specifically, we look at what happens when (i) there are multiple mediators, some of which may be omitted from estimation, (ii) the assumption of strict exogeneity of M is violated, and (iii) the treatment is completely defined by the mediator.

The remainder of this paper is organized as follows. In section 2, we introduce the necessary point-identification assumptions of the front-door criterion and a "how-to" for economists wishing to broaden their empirical toolkit by incorporating the front-door criterion. Section 3 presents two empirical illustrations, one using simulated data and the other using real-world data. In section 4, we explore departures from some of the assumptions underpinning the FDC. We conclude in section 5 by offering practical recommendations for using the FDC in empirical research.

2 The Front-Door Criterion: Identification and Estimation

We begin this section by formally introducing the front-door criterion (FDC) estimand. We first present the necessary identification assumptions noted by Pearl (1995, 2000), and demonstrate how these assumptions recover an unbiased estimate of the ATE of X on Y with the presence of unobserved confounders U, as shown in Figure I above. We then offer our first contribution by explaining how to use the FDC in a linear regression context.

2.1 Identification

We are interested in estimating the ATE of X on Y in Figure I above. Recall that with observational data, estimating the ATE is complicated by the presence of unobserved confounders, U, which give rise to the identification problem. Given the validity of a number of identifying assumptions, however, the FDC approach pictured in Figure I allows for an unbiased point estimate of the ATE of X on Y.

As discussed in Pearl (1995, 2000), the FDC requires that there exists a variable M which satisfies the following assumptions relative to X and Y:[8]

Assumption 1. *The only way in which X influences Y is through M.*

In Figure I, this means that there should be no arrows bypassing M between X and Y. In Pearl's terminology, M should intercept all directed paths from X to Y.

Assumption 2. *The relationship between X and M is not confounded by unobserved variables.*

That is, the coefficient $\gamma = P(M|X)$ in Figure I is identified. In Pearl's terminology, there can be no back-door path between X and M.

Assumption 3. *Conditional on X, the relationship between M and Y is not confounded by unobserved variables.*

That is, the coefficient $\delta = P(Y|M, X)$ in Figure I is identified. In Pearl's terminology, every back-door path between M and Y has to be blocked by X.[9]

As in Pearl (1995), we derive the FDC estimand in three steps, aiming to compute $P(Y|do(X))$ with observable variables,[10] where $P(Y|do(X))$ represents the ATE of X on Y.[11] As shown in Figure I, observing $do(X)$ is complicated by the presence of the unobserved confounder U. Therefore, our goal here is to restate $P(Y|do(X))$ using only the observed variables M, X, and Y while exploiting Assumptions 1 through 3.

[8] In this discussion we consider M, X, and Y to be binary variables. This is only for ease of exposition in introducing the FDC assumptions. In practice, M, X, and Y can each be continuous variables and M can be a vector of variables.

[9] This assumption is conceptually equivalent to the ignorability assumption required for matching estimators to estimate treatment effects (Rosenbaum and Rubin 1983). Recall that a nonrandom treatment can be considered ignorable if confounders are removed. Effectively, ignorability is upheld if the treatment is "as good as random," conditional on confounders. In this sense, it is similar to exogeneity of treatment.

[10] For readers who are not familiar with the $do(\cdot)$ notation, Pearl (1995) introduces $do(\cdot)$ as shorthand for an intervention that sets the variable in parentheses to a specific value. Thus, $P(Y|do(X = x))$ denotes the probability of Y when X is set equal to x by researcher intervention, or when X is manipulated and everything else is held constant (Haavelmo 1943; Strotz and Wold 1960; Heckman et al. 2013). Although $do(X)$ does not necessarily imply random assignment, $do(X)$ should be read as "X is (as good as) randomly assigned."

[11] This should be contrasted with $P(Y|X)$, which may not represent the ATE of X on Y due to the presence of the unobserved confounder U.

The first step is to compute $P(M|do(X))$. Under Assumption 2, the lack of a back-door path between X and M implies the relationship between X and M is identified. When that assumption holds, we can write

$$P(M|do(X)) = P(M|X), \tag{1}$$

given that in this case, the unobserved confounder U affecting X but not M makes the two sides of Equation 1 equivalent.

The second step is to compute $P(Y|do(M))$. Here we cannot set $do(M) = M$ because there is a back-door path from M to Y via X. To block this path we use / Assumption 3. Conditional on X, the relationship between M and Y is not confounded by unobserved variables. In that case, by controlling for and summing over all observations, indexed by i, X_i of X, we can write

$$P(Y|do(M)) = \sum_X P(Y|X, do(M)) \times P(X|do(M)) \tag{2}$$

where the right-hand-side of Equation 2 involves two expressions involving $do(M)$. The second term on the right-hand-side of Equation 2 can be reduced to $P(X)$ because, as stated by Assumption 1, the only way in which X influences Y is through M.[12] The first term on the right-hand-side of Equation 2 can be expressed as $P(Y|X, M)$ because, as stated by Assumption 3, conditional on X, the relationship between M and Y is not confounded. Therefore, we can write

$$P(Y|do(M)) = \sum_X P(Y|X, M) \times P(X). \tag{3}$$

The third and last step is to combine the two effect estimates, $P(M|do(X))$ from Equation 1 and $P(Y|do(M))$ from Equation 2, in order to compute $P(Y|do(X))$—the ATE of X

[12]Since M is a descendent of X in Figure I, any exogenous variation in M will not influence X.

on Y.

To start with, we express $P(Y|do(X))$ in terms of $do(X)$ by controlling for and summing over all observations, indexed by i, M_i of M. This allows us to write

$$P(Y|do(X)) = \sum_M P(Y|M, do(X)) \times P(M|do(X)). \tag{4}$$

Assumption 3 allows us to rewrite M as $do(M)$ in the first term on the right-hand-side of Equation 4. Since, conditional on X, the relationship between M and Y is not confounded, the variation in M is conditionally exogenous. Additionally, as stated by Assumption 1, the only way in which X influences Y is through M, and so we can remove $do(X)$ from the first term on the right-hand side of Equation 4. Said differently, M should have no effect on X, because X causes M and not vice versa in Figure I. Therefore, we can rewrite the first term on the right-hand side of Equation 4 as

$$P(Y|M, do(X)) = P(Y|do(M), do(X)) = P(Y|do(M)). \tag{5}$$

Recall that Equation 3 states that $P(Y|do(M)) = \sum_X P(Y|X, M) \times P(X)$ and Equation 1 states that $P(M|do(X)) = P(M|X)$. Therefore, plugging Equation 3 into Equations 4 and 5, and plugging Equation 1 into Equation 4 gives us the FDC estimand as originally derived by Pearl (1995). That estimand is such that

$$P(Y|do(X)) = \sum_M P(M|X) \times \sum_{X'} P(Y|X', M) \times P(X'). \tag{6}$$

In later writings on the FDC, Pearl (2000) discusses an additional condition for identification, a data requirement which can be directly be verified, and which thus need not be assumed. That condition states that no matter what the value of the mediator M is for unit i, that unit has to have a nonzero probability of getting treated, and thus the mediator M cannot be entirely defined by the treatment X. That is, $P(X_i|M_i) > 0$. In Pearl's canonical

example of the relationship between smoking X and lung cancer Y, this condition implies that the amount of tar in the lungs of smokers M must be the the result not only of smoking, but also of other factors (e.g., exposure to environmental pollutants), and that tar be absent from the lungs of some smokers (say, because of an extremely efficient tar-rejecting mechanism). We will discuss this condition in more detail in Section 4.

2.2 Estimation

We now discuss how to empirically estimate treatment effects using the FDC. As stated above, our goal is to estimate the ATE of X on Y in Figure I, which is defined as $P(Y|do(X))$ and is not equivalent to $P(Y|X)$ because of the presence of unobserved confounders U. When the necessary identification assumptions for the FDC hold, we can estimate the ATE is by using the following approach. Let

$$M_i = \kappa + \gamma X_i + \omega_i \tag{7}$$

and

$$Y_i = \lambda + \delta M_i + \phi X_i + \nu_i. \tag{8}$$

In Equation 7, following Assumption 2 which states that the only way in which X influences Y is through M, the relationship between X and M is identified, since $Cov(X, \omega) = 0$. In Equation 8, Y_i is the outcome variable, which is related to X_i only through M_i. In this case, following Assumptions 1 and 3 (which together imply that the only way X influences Y is through M) and conditional on X, the relationship between M and Y is not confounded, since $Cov(M, \nu) = 0$. Therefore, estimating Equations 7 and 8 and multiplying coefficient estimates $\hat{\delta}$ and $\hat{\gamma}$ by each other estimates β, the ATE of X on Y.

At this point, it is important to note a few things for clarity. First, we focus here on the context of linear regression because linear regression is the approach favored by the majority of applied economists. We note, however, that although we have written

Equations 7 and 8 as linear equations, directed acyclic graphs such as the one in Figure I impose no such linear relationships on their constituent variables, nor do they impose that the relationships be parametric.[13] Therefore the FDC is nonparametrically identifiable, and linear regression is but one way to estimate treatment effects using the FDC.

Second, the necessary identification Assumptions 1 through 3 lead to $Cov(X, \omega) = 0$ in Equation 7 and $Cov(M, \nu) = 0$ in Equation 8. This allows for the unbiased estimation of δ and γ, and via multiplication, β, the ATE of X on Y. Given a additional conditional ignorability assumption (Rosenbaum and Rubin 1983), these same conditions can be achieved by conditioning on a vector of control variables. This is akin to conditional excludability in instrumental variable estimation (Angrist and Kruger 1995), and we will illustrate this in Section 3.

Finally, in our applications we estimate the FDC using a seemingly unrelated regressions (SUR) framework (Zellner 1962). Although the SUR framework is not explicitly necessary to estimate treatment effects using the FDC, it does have some useful features, such as ease of computation.

3 Empirical Illustration

We first show empirical results using simulated data. We then demonstrate the first empirical application of the FDC to observational data wherein the required assumptions plausibly hold for point-identification of the average treatment effect. Additionally, in Appendix 3, we replicate the experimental estimates of Beaman et al. (2013) using the FDC approach.

[13] See Morgan and Winship (2015) for an introduction to directed acyclic graphs as they are used in causal inference, and see Pearl (2000) for an in-depth treatment.

3.1 Simulation Results

Our simulation setup is as follows. Let $U_i \sim N(0,1)$, $Z_i \sim U(0,1)$, $\epsilon_{Xi} \sim N(0,1)$, $\epsilon_{Mi} \sim N(0,1)$, and $\epsilon_{Yi} \sim N(0,1)$ for a sample size of $N = 100,000$ observations.[14] Then, let

$$X_i = 0.5U_i + \epsilon_{Xi}, \tag{9}$$

$$M_i = Z_i X_i + \epsilon_{Mi}, \tag{10}$$

and

$$Y_i = 0.5M_i + 0.5U_i + \epsilon_{Yi}. \tag{11}$$

This fully satisfies Pearl's (1995, 2000) three identification assumptions for the FDC to be able to estimate the average treatment effect of X on Y, viz. (i) the only way in which X influences Y is through M, (ii) the relationship between M and X is not confounded by U, since U only affects X and not M, and (iii) conditional on X, the relationship between M and Y is not confounded by U. This simulation setup also satisfies Pearl's additional data requirement that $P(X_i|M_i) > 0$. By substituting Equation 10 into Equation 11, it should be immediately obvious to the reader that the true ATE is equal to 0.250 in our simulations.

To show that the FDC estimates the ATE of X on Y, we estimate three specifications. The first specification, which we refer to as our benchmark specification because it generates an unbiased estimate of the ATE by virtue of controlling for the unobserved confounder U, estimates

$$Y_i = \alpha_0 + \beta_0 X_i + \zeta_0 U_i + \epsilon_{0i}, \tag{12}$$

where, because both X_i and U_i are included on the right-hand-side, $E(\hat{\beta}_0) = \beta$, i.e., the true ATE.

[14] We allow Z_i to be a random coefficient here to be consistent with subsequent simulations discussed later in this paper.

TABLE I: Simulation Results—Ideal Case

Variables	Benchmark	Naïve	Front-Door		Direct Effect
	Y	Y	M	Y	Y
	(1)	(2)	(3)	(4)	(5)
Treatment (X)	0.252***	0.454***	0.507***	0.200***	-0.003
	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Mediator (M)	–	–	–	0.502***	0.500***
				(0.003)	(0.003)
Confounder (U)	0.499***	–	–	–	0.501***
	(0.004)				(0.004)
Intercept	-0.004	-0.005	-0.004	-0.003	-0.003
	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)
Estimated ATE	0.252***	0.454***	0.254***		–
	(0.004)	(0.003)	(0.002)		
Observations	100,000	100,000	100,000		100,000

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The front-door equations in columns (3) and (4) are estimated by seemingly unrelated regressions. The standard error for the front-door ATE is estimated by the delta method.

Second, we estimate a naïve specification. The naïve specification differs from the benchmark specification in Equation 12 by failing to control for the presence of the unobserved confounder.

The last specification, which we refer to as our front-door specification, estimates

$$M_i = \kappa_0 + \gamma_0 X_i + \omega_{0i} \tag{13}$$

$$Y_i = \lambda_0 + \delta_0 M_i + \phi_0 X_i + \nu_{0i} \tag{14}$$

where the unobserved confounder U_i does not appear anywhere, but because the necessary assumptions for the FDC to identify the ATE hold, $E(\hat{\gamma}_0 \cdot \hat{\delta}_0) = \beta$, i.e., the true ATE.

Column 1 of Table I shows estimation results for Equation 12, our benchmark specification. Column 2 shows estimation results for our naïve specification. Columns 3 and 4 show estimation results respectively for the front-door specification in Equations 13 and

14, respectively. The line labeled "Estimated ATE" shows estimates of the ATE for each of those three specifications. Unsurprisingly, the estimates of the ATE in columns 1 and 2 differ markedly, as the former controls for U_i but the latter does not: $\hat{\beta}$ is equal to 0.252 in the benchmark case, but it is near double that at 0.454 in the naïve case.

Given the derivations above, it should also be unsurprising that the ATE estimate generated by multiplying the coefficient on treatment in column 3 by the coefficient on the mediator in column 4 is equal to 0.254. Assuming the ATE in column 1 is not correlated with the ATE computed from columns 3 and 4, the benchmark and front-door ATEs are statistically identical. In both cases, the estimated ATE is not statistically different from its true value of 0.250.

Column 5 in Table I serves an illustrative purpose. It shows that conditional on the mediator (M) and the unobserved confounder (U), the coefficient on the treatment (X) is statistically indistinguishable from zero. This result highlights the "no direct effect" assumption that is implied by Assumptions 1 through 3.

Finally, we slightly alter our simulation's data generating process to show that it is possible to use the FDC approach when the mediator M is not strictly, but only conditionally exogenous to the relationship between X and Y. In this setup, we add F_i, an observed confounder that captures both selection into treatment and into the mechanism, while also affecting outcome. Let $F_i \sim N(0,1)$. Then, let

$$X_i = 0.5U_i + 0.5F_i + \epsilon_{Xi}, \tag{15}$$

$$M_i = Z_i X_i + 0.3F_i + \epsilon_{Mi}, \tag{16}$$

and

$$Y_i = 0.5M_i + 0.5U_i + 0.15F_i + \epsilon_{Yi}. \tag{17}$$

.

In this case, our benchmark specification controls for both unobserved and observed

TABLE II: Simulation Results—Conditionally Exogenous Mediator

Variables	Benchmark	Naïve	Front-Door	
	Y	Y	M	Y
	(1)	(2)	(3)	(4)
Treatment (X)	0.251***	0.452***	0.499***	0.200***
	(0.004)	(0.003)	(0.003)	(0.003)
Mediator (M)	–	–	–	0.504***
				(0.003)
Observed Confounder (F)	0.303***	0.204***	0.304***	0.052***
	(0.004)	(0.004)	(0.004)	(0.004)
Unobserved Confounder (U)	0.499***	–	–	–
	(0.004)			
Intercept	-0.004	-0.005	-0.004	-0.003
	(0.004)	(0.004)	(0.003)	(0.004)
Estimated ATE	0.251***	0.452***	0.251***	
	(0.004)	(0.003)	(0.002)	
Observations	100,000	100,000	100,000	

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The front-door equations in columns (3) and (4) are estimated by seemingly unrelated regressions. The standard error for the front-door ATE is estimated by the delta method.

confounders. Thus, it estimates

$$Y_i = \alpha_1 + \beta_1 X_i + \rho_1 F_i + \zeta_1 U_i + \epsilon_{1i}, \tag{18}$$

where the right-hand-side of Equation 18 includes both the observed confounder F_i and the unobserved confounder U_i. Similar to the previous simulation, the naïve specification differs from the benchmark specification by failing to control for the presence of the unobserved confounder, U_i.

Finally, the FDC estimates

$$M_i = \kappa_1 + \gamma_1 X_i + \tau_1 F_i + \omega_{1i} \tag{19}$$

$$Y_i = \lambda_1 + \delta_1 M_i + \phi_1 X_i + \pi_1 F_i + \nu_{1i} \tag{20}$$

where both Equations 19 and 20 include F_i on the right hand side.

Table II presents results of this simulation with 100,000 observations. Column 1 displays results of the benchmark regression, with the coefficient on X indicating the "true" ATE of 0.251. Unsurprisingly the naïve specification overestimates the ATE and the FDC estimation approach yields an unbiased estimate of the ATE. These results highlight that, given an additional conditional ignorability assumption (Rosenbaum and Rubin 1983), the FDC approach can point-identify treatment effects by conditioning on a vector of control variables.

3.2 Real-World Application: Ride Sharing and Tipping Behavior

Using publicly available data on over 890,000 Uber and Lyft rides in Chicago between June 30 and September 30, 2019, we use the FDC to estimate the ATE of authorizing a shared ride on tipping at both the extensive (i.e., whether the passenger tips) and intensive margins (i.e., how much the passenger tips). We find that naïve regressions overestimate the magnitude of the ATE of authorizing sharing on both tipping margins because of selection into treatment. We show that the necessary conditions for the FDC to yield a consistent ATE apply in this scenario after conditioning on relevant observed variables. Although sharing-authorized rides are not determined exogenously, whether a passenger actually ends up sharing a ride with another is plausibly exogenous conditional on several observable factors.

The only path through which sharing authorization (i.e., X) will affect tipping (i.e., Y) is through whether a ride is actually shared (i.e., M). This mediator is plausibly relevant because tipping variation in ride-hailing and taxi settings can be affected by demand-side factors such as rider experience, mood, and social preferences (Chandar et al. 2019). For example, a passenger's experience or mood may be worsened by sharing a car with a stranger or if a driver takes additional time to drop off or pick up another passenger. This may be especially true if passengers initially authorized sharing hoping that their ride

would fall into the share of sharing-authorized rides that do not overlap with another passenger's trip.[15] Tipping is also motivated by a desire to avoid unpleasant feelings of guilt and embarrassment (Azar 2020). However, the guilt associated with not tipping or tipping a low amount may decline when a passenger knows another rider may tip—behavior which is reminiscent of the free-rider problem (Boyes et al. 2006).

3.2.1 Background

Examining 40 million UberX (i.e., solo) rides during the summer of 2017, Chandar et al. (2019) find that "demand-side" factors that capture an individual consumer's propensity to tip explain more of the variation in tipping than "supply-side" factors such as driver or ride quality. By examining only solo rides, however, Chandar et al. (2019) omit a key determinant of tipping: whether a passenger opts to share a ride. After opening her Uber app, a passenger can select either UberX or UberPool, with the latter option allowing one's ride to overlap with the rides of more than one passenger in the same vehicle while enjoying a discount on the fare price (Hemel 2017). Lyft offers a similar service called Lyft Line. Fares for both services are charged up front and are calculated according to the probability a given passenger ends up sharing a hailed vehicle with another stranger. Discounts from the single-passenger "base fare" are set for sharing-authorized rides such that rides more likely to overlap with another passenger's trip are cheaper relative to the base fare. Notably, the higher the stated fare for a single-passenger ride, the more likely a passenger is to authorize sharing (Wu et al. 2018).

Using an earlier version of the same data we use here, data-analytics firm CompassRed (2019) finds that when riders opt to share rides with another passenger using the Lyft Line or UberPool services, they are less likely to tip. In the context of the potential outcomes model (Rubin 2005), this finding is problematic because it fails to account for selection into treatment. Customers who are frugal are both less likely to tip and more likely to

[15] In our data, roughly 40% of sharing authorized rides end up not being shared.

authorize sharing, enticed by lower fares. To effectively infer the ATE of authorizing a shared ride on tipping, one must deal with the endogeneity associated with selection of people with a lower propensity to tip into authorizing shared rides. An unbiased ATE would capture the difference in tipping if UberPool or Lyft Line rides were randomly assigned across all passengers, no matter their proclivity to tip.

3.2.2 Data

Our data include 890,000 dedicated (i.e. standard, "single-transaction" UberX and Lyft rides) and sharing-authorized Uber and Lyft rides taken within the city limits of Chicago from June 30 to September 30, 2019. The data come from the Chicago Department of Business Affairs and Consumer Protection's Transportation Network Providers Data Portal.[16] Each observation represents a single transaction on either app. These data show whether the passenger authorized a shared ride (i.e., X), whether the passenger actually shared a ride with another paying customer (i.e., M),[17] and the passenger's tipping behavior at both the extensive and intensive margins (i.e, Y).

These data provide the base fare (rounded to the nearest \$2.50) and tip amount (rounded to the nearest \$1.00).[18] For the extensive margin of tipping, our dependent variable is a binary variable capturing whether a passenger tips. For the intensive margin, we use the observed tip value.[19] Additionally, we generate several sets of fixed effects from observed time and geographic indicators, including for each origin–destination pair of community

[16]The data are the first and only publicly available data on transportation network company trips and have been collected since November 2018. They can be downloaded via the City of Chicago's website.

[17]The data show the number of overlapping sharing-authorized rides a given ride occurred within. Specifically, this field counts how many individual passengers were transported between any two points in time during which the car was occupied by passengers. Any number over one indicates that a ride was shared with at least one other passenger.

[18]We discuss the measurement error introduced by these respective rounding schemes below, when interpreting our results. We also drop observations with fare level under \$2.50 and over \$50 to analyze a reasonable range of fares.

[19]To account for the high number of zero observations (indicating that a passenger did not tip), and because we would ideally want to take the logarithm of tip value, we apply the inverse hyperbolic sine (i.e., arcsinh) transformation, a log-like transformation which allows to keep the zero-valued observations, before calculating elasticities (see derivations in Bellemare and Wichman 2019, and see Card et al. 2020 for an application).

TABLE III: Summary Statistics

		Fare ($)		Tip ($)		Tipped (Dummy)		Observations	
	Ride Type	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	N	% Total
Full Sample	Dedicated	10.842	(6.884)	0.659	(1.602)	0.215	(0.411)	750,883	84.2%
	Sharing Authorized	9.055	(5.401)	0.208	(0.815)	0.092	(0.289)	140,446	15.8%
Sharing Authorized	Shared	9.332	(5.628)	0.207	(0.791)	0.092	(0.289)	88,372	62.9%
	Not Shared	8.583	(4.957)	0.209	(0.853)	0.091	(0.288)	52,074	37.1%

areas in Chicago.[20] Summary statistics are provided in Table III.

Though our M variable (i.e., whether a passenger who authorized a shared ride actually get to share a ride) is not strictly exogenous to either our X (i.e., whether a passenger authorizes a shared ride) and Y (i.e., tipping behavior) variables, we argue that it is *conditionally* exogenous to both those variables. Indeed, we first condition on a ride's fare level, exploiting app algorithms that set fares according to the likelihood a ride is ultimately shared. This helps control for the propensity that any given ride is shared. Although this variable does help control for endogenous factors embedded in the app's algorithms, the rounding of these data to the nearest $2.50 makes conditioning on fare level less precise. Second, to further condition away potential endogeneity between the likelihood a (sharing-authorized) ride is actually shared on the one hand and tipping on the other hand, we control for date, hour, day of the week–hour, and origin–destination fixed effects. Between controlling for fare level and those several layers of fixed effects, we plausibly uphold the assumption of ignorability (Rosenbaum and Rubin 1983). In other words, although M may not be strictly exogenous to X and Y, our controls ensure that M is plausibly exogenous to X and Y in this application.[21]

Our estimation strategy consists of estimating the following equations.

$$\text{Naïve: } Y_i = \alpha_2 + \beta_2 X_i + \rho_2 F_i + \boldsymbol{T_i'\chi_2} + \boldsymbol{G_i'\theta_2} + \epsilon_{2i} \tag{21}$$

[20] Chicago is divided into 77 community areas for policy, planning, and statistical purposes. The data show the origin and destination community areas, which we use to develop the origin–destination pair fixed effects.

[21] We include the same exact set of controls in both stages of FDC estimation. This is because the exact same controls are necessary to uphold conditional exogeneity of both X and M. In other applications, it may not be necessary to include identical sets of controls if conditional exogeneity at different stages of estimation can be upheld through conditioning on different sets of observables.

$$\text{FDC First Stage: } M_i = \kappa_2 + \gamma_2 X_i + \tau_2 F_i + \boldsymbol{T_i'\xi_2} + \boldsymbol{G_i'\sigma_2} + \omega_{2i} \tag{22}$$

$$\text{FDC Second Stage: } Y_i = \lambda_2 + \delta_2 M_i + \phi_2 X_i + \pi_2 F_i + \boldsymbol{T_i'\psi_2} + \boldsymbol{G_i'\iota_2} G_i + \nu_{2i} \tag{23}$$

where Y now represents tipping at either the extensive or intensive margin, F_i is the fare level, $\boldsymbol{T_i}$ is a vector of time fixed effects (i.e., date, hour, and day of the week–hour fixed effects), and $\boldsymbol{G_i}$ is a vector of geographic fixed effects (i.e., origin-destination pairs). Additionally, X_i is our treatment variable, which indicates whether a passenger authorized ride-sharing, and M_i indicates whether the ride was actually shared with another passenger.

We estimate the two FDC equations by seemingly unrelated regression (Zellner 1962) to account for the potential correlation between the two equations 22 and 23. To recover the ATE of X on Y estimated by the FDC, we simply multiply the coefficient estimates $\hat{\gamma_2}$ and $\hat{\delta_2}$ by each other. Because the ATE is a nonlinear combination of coefficients, standard errors for the ATE estimated by the FDC are obtained using the delta method.

In this empirical application, the identifying assumptions follow those discussed in Section 2. First, the only way in which X influences Y is through M. This assumption is supported by the fact that the only way authorizing a shared a ride X can ever influence tipping behavior Y is if the passenger actually gets to share a ride M. Second, $Cov(X, \omega_2) = 0$. This assumption is supported by the fact that M is determined by X and the embedded app algorithm. Third, $Cov(M, \nu_2) = 0$. Conditional on X, F, $\boldsymbol{T}$, and $\boldsymbol{G}$ we argue that this is a valid assumption. Finally, $P(X_i|M_i) > 0$, which is valid because if $M_i = 1$ then $X_i = 1$ and if $M_i = 0$ then $X_i = 1$ or $X = 0$.

3.2.3 Results

Table IV shows results for tipping at the extensive margin. In this case, the naïve specification estimates that authorizing sharing reduces the probability a rider will tip by 5.8%. The FDC, however, estimates that authorizing sharing reduces tipping probabil-

TABLE IV: The Effect of Authorizing Sharing on Tipping at the Extensive Margin

Variables	Naïve	Front-Door	
	Tipped	Shared Trip	Tipped
	(1)	(2)	(3)
Sharing Authorized (X)	-0.073***	0.621***	-0.067***
	(0.001)	(0.001)	(0.001)
Shared Trip (M)	–	–	-0.009***
			(0.002)
Intercept	0.114***	0.0882***	0.115***
	(0.006)	(0.003)	(0.006)
Estimated ATE	-0.073***	-0.006***	
	(0.001)	(0.001)	
Elasticity	-5.8%***	-0.5%***	
	(0.001)	(0.001)	
Observations	891,329	891,329	
R-squared	0.051	0.614	0.051

Notes: Both specifications control for fare level, date, hour, day of the week–hour, and origin–destination fixed effects. FDC specification estimated using seemingly unrelated regression. Standard errors for the FDC ATE and ATE elasticity computed using the delta method. *** p<0.01, ** p<0.05, * p<0.1.

ity by only 0.5%, an ATE a full order of magnitude smaller than the naïve estimate. The difference between the two ATE estimates suggests that much of the naïve specification simply captures endogeneity of selection into treatment.

Table V shows results for tipping at the intensive margin. These results mirror those for tipping at the extensive margin. The naïve specification finds that authorizing sharing reduces the tip by about 3.6%. The FDC presents a much lower, yet still statistically significant effect. The ATE calculated using the FDC finds that authorizing sharing reduces tip payments by about 0.2%, an effect that is less than a tenth of the magnitude of the naïve ATE estimate.

This application demonstrates the usefulness of the FDC. By exploiting the conditional exogeneity of ride sharing, we can estimate the ATE of authorizing a shared ride on tipping. As it turns out, the FDC estimates lower ATEs than those estimated by the naïve specification, at both the intensive and extensive margins of tipping. This result

TABLE V: The Effect of Authorizing Sharing on Tipping at the Intensive Margin

Variables	Naïve	Front-Door	
	arcsinh(Tip)	Shared Trip	arcsinh(Tip)
	(1)	(2)	(3)
Sharing Authorized (X)	-0.127***	0.621***	-0.119***
	(0.002)	(0.001)	(0.002)
Shared Trip (M)	–	–	-0.013***
			(0.003)
Intercept	0.113***	0.0882***	0.114***
	(0.010)	(0.003)	(0.010)
Estimated ATE	-0.127***	-0.008***	
	(0.002)	(0.001)	
Elasticity	-3.6%***	-0.2%***	
	(0.001)	(0.001)	
Observations	891,329	891,329	
R-squared	0.084	0.614	0.084

Notes: Both specifications control for fare level, date, hour, day of the week–hour, and origin–destination fixed effects. FDC specification estimated using seemingly unrelated regression. Standard errors for the FDC ATE and ATE elasticity computed using the delta method. *** p<0.01, ** p<0.05, * p<0.1.

suggests that if one had conducted a randomized controlled trial wherein passengers are randomly assigned to either a dedicated (i.e., single-passenger) or sharing-authorized ride in Chicago between June 30 and September 30, 2019, one would have estimated effects of sharing-authorized rides on tipping that are similar to ours.

From a business strategy perspective, this suggests that *ceteris paribus*, making all rides sharing-authorized as the default setting out of which passengers would have to opt out could lead to higher tips for those firms' drivers by exploiting status quo bias (Kahneman et al. 1991; Fernandez and Rodrik 1991). Indeed, if all passengers are *de facto* assigned to a sharing-authorized ride out of which they have to opt out should they prefer to ride alone, some passengers who would otherwise have not chosen to share a ride will remain in the sharing-authorized category because of the status quo bias. Our results indicate that this could lead to a greater likelihood that passengers will tip and tip larger amounts

on shared rides than otherwise.[22] Furthermore, an increase in the number of shared rides is likely to lead in decreased costs for ride-hailing firms via ride consolidation. Ultimately, this could lead to increased profits for those firms without limiting passenger agency, though this claim is obviously speculative given that we do not observe the cost structure of ride-hailing firms.

The data have a few weaknesses. First, the data do not differentiate between Uber and Lyft rides. Though it is likely that Uber and Lyft employ different algorithms for setting fares once a rider opts to authorize sharing, we are confident that our strategy of conditioning on fixed effects adequately takes care of this issue.

Additionally, we do not observe the exact tip or fare payments, observing rounded values instead.[23] This means that in columns 2 and 3 of Table V, we are dealing with two sources of classical measurement error. The first is classical measurement error in fare level,[24] which is a control variable in both columns 2 and 3. We are not worried about this source of measurement error because it merely biases the coefficient on fare level—a control variable whose coefficient is not directly of interest in our analysis—toward zero. The second source of measurement error is classical measurement error in tipping amount, i.e., the dependent variable, in column 3. This is in principle more problematic because classical measurement error in the dependent variable leads to less precise estimates. Though this would be worrisome in a small sample because it could lead to a type II error (i.e., we would fail to reject the null hypothesis that the coefficient on M is equal to zero), this is not an issue in our a sample of over 890,000 observations—we indeed reject the null hypothesis that the coefficient on M in column 3 is equal to zero.

[22] Alternatively, firms could mandate a minimum tip payment for both sharing-authorized and dedicated rides.

[23] This mainly challenges the calculation of the effect of authorizing sharing on tipping percentage, a potentially interesting causal relationship which we do not explore because the dependent variable (i.e., tipping percentage) would have to be calculated on the basis of two variables measured with error.

[24] For example, because fares are rounded at the nearest \$2.50 in the data, a reported \$15 ride's true fare could lie anywhere in the $(\$13.75, \$16.25)$ interval.

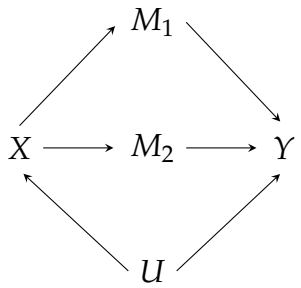
4 Departures from the Ideal Case

Having discussed how identify and estimate ATEs with the FDC in section 2, and having illustrated the use of the FDC to estimate ATEs using both simulation and real-world data in section 3, we now turn to investigate what happens when some of the assumptions required for the FDC to identify an ATE fail to hold. To do so, we look in turn at what happens with multiple mediators, when the mediator is no longer strictly exogenous, and when the treatment is totally defined by the mediator.

4.1 Multiple Mediators

Pearl's (1995, 2000) canonical treatment of the front-door criterion assumes that M is a single variable, and not a vector of mediator variables. Consequently, in the empirical examples in section 3, we considered cases where the mediators, M, is defined by a single variable rather than a vector. In this sub-section we consider how to implement a case where we have multiple mediators.

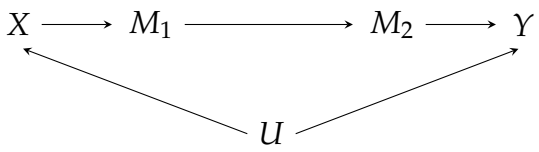
FIGURE II: Multiple Mediators—Case 1



There are two basic cases in which we can imagine multiple mediators. Of course, one can imagine more complicated cases that combine these two basic cases. For illustrative purposes, however, we will examine these two cases separately. In the first case, as shown in Figure II, the multiple mediators are independent from each other. Specifically, a path flows from X to both M_1 and M_2, and additionally, a path flows from both M_1 and M_2 to Y. In this case, M_1 and M_2 together intercept all directed paths from X to Y and meet

the requirement Assumption 1.[25] By simply examining Figure II it is clear that omitting either M_1 or M_2 from the estimation will violate Assumption 1, since the single mediator does not intercept all directed paths from X to Y.

FIGURE III: Multiple Mediators—Case 2



In the second case, as shown in Figure III, the multiple mediators both lie on the same path between X and Y. Specifically, a path flows from X to M_1, from M_1 to M_2, and finally from M_2 to Y. In this case, either M_1 or M_2 intercept all directed paths from X to Y and meet the requirement of Assumption 1. In contrast to the previous case, omitting either M_1 or M_2 from the estimation will not violate Assumption 1, since both mediators individually intercept all directed paths from X to Y. Therefore the FDC approach will recover the ATE when using either only M_1, only M_2, or both M_1 and M_2 as mediators in the FDC estimation. This point should be obvious based on conceptual reasoning, but a simulation showing this result can be found in Appendix 2.

We now show simulation results that demonstrate the consequences of multiple mediators of the sort illustrated in Figure II, where multiple mediators lie on different paths from X to Y.

Our simulation setup is as follows. Let $U_i \sim N(0,1)$, $\epsilon_{Xi} \sim N(0,1)$, $Z_{1i} \sim U(0,1)$, $Z_{2i} \sim U(0,1)$, $\epsilon_{M1i} \sim N(0,1)$, $\epsilon_{M2i} \sim N(0,1)$, and $\epsilon_{Yi} \sim N(0,1)$ for a sample size of $N = 100,000$ observations. Then, let

$$X_i = 0.5U_i + \epsilon_{Xi}, \tag{24}$$

$$M_{1i} = Z_{1i}X_i + \epsilon_{M1i}, \tag{25}$$

[25]This case, where M_1 and M_2 together intercept all directed paths from X to Y is similar to the surrogate index of Athey et al. (2019).

$$M_{2i} = Z_{2i}X_i + \epsilon_{M2i}, \tag{26}$$

and

$$Y_i = 0.5M_{1i} + 0.5M_{2i} + 0.5U_i + \epsilon_{Yi}. \tag{27}$$

As illustrated in Figure II, this fully satisfies Pearl's (1995, 2000) three assumptions for the FDC to be able to estimate the average treatment effect of X on Y. By substituting Equations 25 and 26 into Equation 27, it should be obvious to the reader that the true ATE is equal to 0.500 in our simulations.

Similar to the previous simulation analysis, we estimate several specifications. The first baseline specification estimates

$$Y = \alpha_3 + \beta_3 X_i + \zeta_3 U_i + \epsilon_{3i}, \tag{28}$$

where, because both X and U are included on the right-hand-side, $E(\hat{\beta}_1) = \beta$, i.e., the true ATE.

The second specification estimates

$$M_{1i} = \kappa_3 + \gamma_3 X_i + \omega_{3i}, \tag{29}$$

$$M_{2i} = \pi_3 + \rho_3 X_i + \eta_{3i}, \text{ and} \tag{30}$$

$$Y_i = \lambda_3 + \delta_3 M_{1i} + \tau_3 M_{2i} + \phi_3 X_i + \nu_{3i}, \tag{31}$$

where the unobserved confounder U does not appear anywhere. The small difference in the case of multiple independent mediators is the true ATE is calculated by adding two products together, $E[(\hat{\gamma}_3 \cdot \hat{\delta}_3) + (\hat{\rho}_3 \cdot \hat{\tau}_3)] = \beta$.

Column 1 of Table VI shows our benchmark estimation results for Equation 28. Column 2 shows estimation results for the naïve version of Equation 28 which omits the unobserved confounder U. Columns 3, 4, and 5 show FDC estimation results using the

specification outlined in Equations 29 to 31, respectively. Again, the estimates of the ATE in columns 1 and 2 are quite different. While the ATE estimate is equal to 0.501 in the benchmark case, it is much larger, at 0.703, in the naïve case.

Given the derivations above, it should be unsurprising that the FDC approach accurately estimates the ATE. The FDC approach first multiplies the coefficient on X in column 3 by the coefficient on M_1 in column 5. Next, the FDC approach multiplies the coefficient on X in column 4 by the coefficient on M_2 in column 5. Finally, these two products are summed to estimate the ATE. Assuming the ATE in column 1 is not correlated with the ATE computed from columns 3 through 5, the two ATEs are statistically identical. In both cases, the estimated ATE is not statistically different from its true value of 0.500. Finally, in column 6, the direct effect of treatment conditional on M_1, M_2, and U is statistically indistinguishable from zero. More interesting, however, is investigating and interpreting estimates when we erroneously omit one of the mediators (say, for example, M_2) from the FDC estimation. In this case, we no longer can correctly assume no "direct effect" of X on Y since there is a directed path independent of M_1 via M_2. This violates Assumption 1 above. When we omit M_2 from the FDC estimation the estimated ATE (shown in columns 7 and 8) is 0.246, considerably smaller than the true ATE. Column 9 shows that the "direct effect" is 0.254.

The foregoing shows the consequences of omitting a mediator when using the FDC approach to estimate the ATE. With that said, the effect estimated in the biased FDC estimation in columns 7 and 8 of Table VI can be interpreted as the"indirect effect" of X on Y via M_1 and independent of M_2 (Imai et al. 2010, Acharya et al. 2016). In the literature on causal mediation analysis, the total causal effect is framed as the aggregation of both the direct and indirect effects (Imai et al. 2011). We call this effect, estimated using the FDC approach, the mediated average treatment effect (MATE).

A very common approach for estimating indirect or mediating effects is to simply condition on potential mediating variables (Acharya et al. 2016). Despite the popu-

TABLE VI: Simulation Results—Multiple Mediators, Case 1

	Benchmark	Naïve	Front-Door			Direct Effect	Biased Front-Door		Direct Effect
Variables	Y	Y	M_1	M_2	Y	Y	M_1	Y	Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment (X)	0.501***	0.703***	0.497***	0.502***	0.204***	0.001	0.497***	0.457***	0.254***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
Mediator (M_1)	–	–	–	–	0.498***	0.500***	–	0.495***	0.496***
					(0.003)	(0.003)		(0.004)	(0.003)
Mediator (M_2)	–	–	–	–	0.499***	0.499***	–	–	–
					(0.003)	(0.003)			
Confounder (U)	0.498***	–	–	–	–	0.501***	–	–	0.500***
	(0.004)					(0.004)			(0.004)
Intercept	-0.002	-0.003	-0.005	0.002	-0.002	-0.004	-0.005	-0.001	0.001
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Estimated ATE	0.501***	0.703***		0.498***		–	0.246***		–
	(0.004)	(0.004)		(0.003)			(0.002)		
Observations	100,000	100,000		100,000		100,000	100,000		100,000

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The front-door equations in columns (3) and (4) are estimated by seemingly unrelated regressions. The standard error for the front-door ATE is estimated by the delta method.

larity of this approach, conditioning on potential mediating variables can lead to biased estimation—specifically in the case when omitted variables are affected by the treatment X and affect both the potential mediating variable M and the outcome Y (see, e.g., Acharya et al. 2016; Imai et al. 2010).[26] If the assumptions in Section 2.1 hold coupled with the ability to relax Assumption 1—that the M intercepts all paths from X to Y—then the FDC approach allows for valid estimation of MATEs. Of course, whether or not the MATE is a parameter of interest for applied researchers will ultimately depend on the specific application and research question.

4.2 Violations of Strict Exogeneity

Together, Assumptions 2 and 3 imply that the mediator M is excludable. More formally, the strict exogeneity of M implies that $P(U|M,X) = P(U|X)$ and $P(Y|X,M,U) = P(Y|M,U)$. In this sub-section, we examine violations of this assumption. Again, we do this with a simulation analysis.

Our simulation setup is the same as in section 3, except that here we allow for the endogeneity of M. Let $U_i \sim N(0,1)$, $Z_i \sim U(0,1)$, $\epsilon_{Xi} \sim N(0,1)$, $\epsilon_{Mi} \sim N(0,1)$, and $\epsilon_{Yi} \sim N(0,1)$ for a sample size of $N = 100,000$ observations. Then, let

$$X_i = 0.5U_i + \epsilon_{Xi}, \tag{32}$$

$$M_i = Z_i X_i + \Gamma U_i + \epsilon_{Mi}, \tag{33}$$

and

$$Y_i = 0.5M_i + 0.5U_i + \epsilon_{Yi}. \tag{34}$$

The critical difference here is that now, when defining M in equation 33, U is included on the right-hand-side. The parameter Γ defines the strength of the relationship between

[26] Also see the discussion of "collider" bias in Morgan and Winship (2015, p.81).

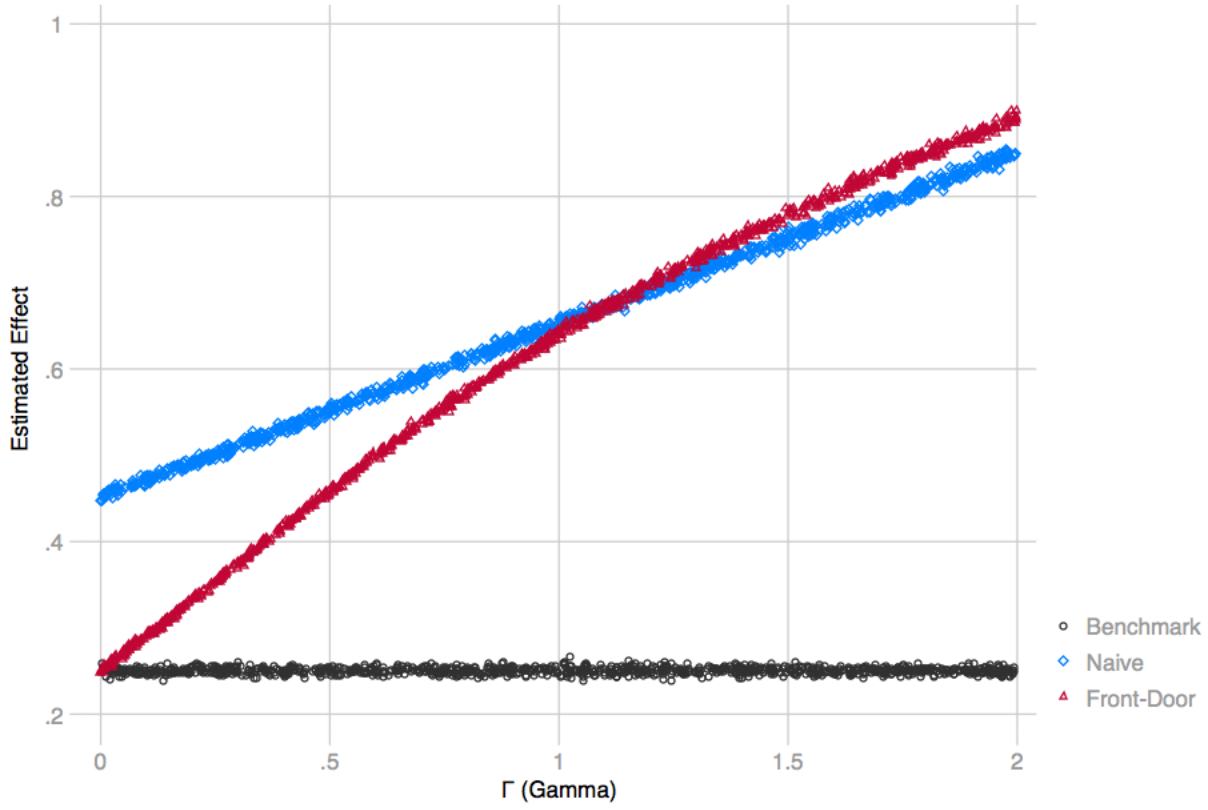
U and M. In this simulation analysis we let $\Gamma \sim U(0, 2)$. By permitting the values of Γ to vary allows the degree of endogeneity in our simulations to vary.

We show these simulation results graphically. Figure IV illustrates how having an endogenous mediator influences the credibility of using the FDC approach to estimate the ATE. This figure shows estimated effects for three estimation approaches. First, the benchmark estimates (black circles), which include the confounder U on the right-hand-side of the regression equation, accurately estimates the true ATE of 0.250. Second, the naïve estimates (blue diamonds), which omits the confounder U from the regression equation, consistently overestimates the ATE. The size of this bias is increasing in the strength of the endogeneity of M. This is because, as Γ increases, the influence of the confounder U in the relationship between X and Y increases. Finally, the front-door estimates (red triangles) are estimated as described in equations 13 and 14 in section 2.

Once again, a few remarks are in order. First, and rather unsurprisingly, it is only when the degree of endogeneity of M is negligible (i.e., when Γ is infinitesimally close to zero) that the FDC approach accurately estimates the ATE. Second, although the FDC approach produces biased ATE estimates, when M is weakly endogenous (i.e., when $\Gamma > 0$ but still relatively small), these estimates are less biased than the naïve estimates. Third, when M is strongly endogenous the FDC approach produces estimates of the ATE that are worse—that is, more biased—than the naïve estimates.

These details lead to an important discussion for applied researchers who may want to implement the FDC approach in their given empirical setting. In many cases, strict exogeneity of M may be debatable. Indeed, outside of an experimental setting, convincingly arguing that $P(U|M, X) = P(U|X)$ and $P(Y|X, M, U) = P(Y|M, U)$ will likely be challenging. That said, however, if applied researchers can convincingly argue that the degree of endogeneity of M is relatively weak—that M is not strictly exogenous but that it is plausibly exogenous (Conley et al., 2012), so to speak—then the FDC approach will produce more reliable estimates of the ATE compared to the naïve approach which con-

FIGURE IV: The Consequences of an Endogenous Mediator



Notes: This figure illustrates simulation results using 1,000 replications from each estimation approach. The vertical axis represents the estimated effect. The horizontal axis represents the Gamma parameter, representing the degree of endogeneity, from equation 32. The benchmark estimates (black circles) all accurately estimate the true ATE of 0.250. The naive estimates are shown in blue diamonds and the front-door estimates are shown in red triangles.

sists in regressing Y on an endogenous X.[27] On the other hand, when the endogeneity of M is obviously relatively strong, using the FDC approach could lead to more bias in estimates of the ATE than the naïve approach. Specifically in our simulation set-up, the FDC estimates begin to become just as biased as the naïve estimates when Γ is equal to one. In the way we have defined our variables, this means that the direct effect of U on M is about twice as strong as the indirect effect of U on M via X. Of course when using real-world data, when we cannot observe U, testing the specific size of these relationships is impossible. In all practical settings, the case for the exogeneity of M will rely on careful reasoning based on the given empirical setting.

4.3 Treatment Totally Defined by the Mediator

Recall that, in addition to the three assumptions in section 2.1 for the FDC to identify the average treatment effect, Pearl (2000) imposed a condition on the data, namely that $P(X_i|M_i) > 0$. This condition requires that for every value of the mediator M, the likelihood that an observation will receive treatment X is nonzero. In other words, the treatment cannot be totally defined by the mediator, and no matter what value the mediator takes, it has to be the case that an observation has a nonzero probability of receiving the treatment.

This requirement is satisfied in our core empirical applications, but it fails to hold in the re-analysis of the Beaman et al. (2013) randomized controlled trial in Appendix 3. In that application, the authors randomly allocated fertilizer to rice farmers in Mali. One group of farmers received the full recommended dose of fertilizer, a second group received half the recommended dose, and a third group received no fertilizer. We defined the treatment (e.g., $X = 1$) if a farmer received any free fertilizer, and zero otherwise. The mediator captured the intensity of treatment (e.g., $M = 1$ if the farmer received the full

[27] Our real-world illustration in the previous section exemplifies a scenario in which a mediator is plausibly exogenous conditional on observed confounders.

dose, $M = 0.5$ if the farmer received the half dose, and $M = 0$ if the farmer received no fertilizer). Therefore, in this case, a farmer who received no fertilizer (e.g., $M = 0$) had a probability of receiving treatment equal to zero (e.g., $X = 0$). Additionally, a farmer who received some fertilizer (e.g., $M = 1$ or $M = 0.5$) had a probability of receiving treatment equal to one (e.g., $X = 1$). Thus, in this application, $P(X_i|M_i) = 0$, which is a violation of Pearl's additional condition (Pearl 2000).[28]

Preliminary work for this paper, however, uncovered the following fact: It is only when there are no unobserved confounders that $P(X_i|M_i) = 0$ is a problem. In such cases, one only need to omit the treatment variable X from estimation in Equation 8 for the method we outline in section 2 to recover the correct ATE. When there are unobserved confounders, the FDC method discussed in section 2 recovers the ATE. We demonstrate these details using simulated data in Table VII, and by showing results using the Beaman et al. (2013) experimental data in Appendix 4.

Our simulation set up here differs slightly from those previously discussed. Let $U_i \sim B(1, 0.5)$, $Z_{Xi} \sim B(1, 0.5)$, $Z_{Mi} \sim B(1, 0.5)$, $\epsilon_{Yi} \sim N(0, 1)$ for a sample size of $N = 100,000$ observations. Then, let

$$X_i = Z_{Xi}U_i \tag{35}$$

$$M_i = Z_{Mi}X_i + X_i \tag{36}$$

$$Y_i = 0.5M_i + 0.5U_i + \epsilon_{Yi} \tag{37}$$

This generates data that characterize a setting where the treatment is totally defined by the mediator. Specifically, the binary treatment X is endogenous to the outcome Y. The mediator M is now binary and is strictly a function of the treatment and can be considered

[28]Though it is obvious how experimental settings may naturally lead to cases where $P(X_i|M_i) = 0$, violations of this assumption are not the exclusive preserve of experimental research designs. Indeed, it is not difficult to imagine observational research designs where only those subjects who select into receiving a given treatment can actually receive that treatment in nonzero amounts. Therefore, this discussion remains relevant for observational research settings.

TABLE VII: Treatment Totally Defined by the Mediator—Endogenous Treatment

	Benchmark	Naïve	Front-Door	
Variables	Y	Y	M	Y
	(1)	(2)	(3)	(4)
Treatment (X)	0.743***	1.078***	1.503***	0.322***
	(0.009)	(0.008)	(0.002)	(0.021)
Mediator (M)	–	–	–	0.503***
				(0.013)
Confounder (U)	0.503***	–	–	–
	(0.008)			
Intercept	-0.005	0.173***	0.000	0.173***
	(0.005)	(0.004)	(0.001)	(0.004)
Estimated ATE	0.743***	1.078***	0.755***	
	(0.006)	(0.008)	(0.019)	
Observations	100,000	100,000	100,000	

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The front-door equations in columns (3) and (4) are estimated by seemingly unrelated regressions. The standard error for the front-door ATE is estimated by the delta method.

akin to treatment intensity. That is, for treated units (i.e., $X = 1$), $M = 1$ or $M = 2$. For untreated units (i.e., $X = 0$), $M = 0$.

Table VII shows that even when the treatment is totally defined by the mediator, the FDC method discussed in section 2 recovers the true ATE when the treatment is endogenous. This does not hold, however, when treatment is exogenous. Table IX, shown in Appendix 4, highlights the fact that when the treatment is exogenous and is totally defined by the mediator, implementing the FDC method as discussed in section 2 will lead to biased estimates of the ATE. Since, treatment is exogenous in this case, however, there is no back-door path from U to X and we do not need to control for treatment in the second stage FDC regression. As it turns out, omitting the treatment variable from the second-stage FDC regression allows for the FDC to recover the ATE.

5 Conclusion

We have focused on the application of Pearl's (1995, 2000) front-door criterion. Because the goal of most research in applied economics nowadays is to answer questions of the form "What is the causal effect of X on Y?," economists should welcome the addition of techniques that allow answering such questions to their empirical toolkit. Yet economists have been reluctant to incorporate the FDC in that toolkit.

We focus here first on explaining how to use the front-door criterion in the context of linear regression, which remains the workhorse of applied economics. Second, we present two empirical examples: one using simulated data, and one relying on observational data on Uber and Lyft rides in Chicago between June 30 and September 30, 2019. Our observational example is, to our knowledge, the first application of the front-door criterion to observational data where the necessary assumptions plausibly hold. Finally, in an effort to help overcome economists' resistance to incorporating the front-door criterion in their empirical toolkit, we look at what happens when the assumptions underpinning the front-door criterion are violated, and what can be done about it in practice.

Our results lead to the following recommendations for applied work:

1. Because the FDC estimand is a nonlinear combination of two estimated coefficients, standard errors can be computed either by the delta method or by bootstrapping. In small samples, bootstrapping should be preferred to the delta method (Davidson and MacKinnon 2004).

2. When the treatment operates through more than one mediator, the average treatment effect is the sum of the mediated average treatment effects (MATEs), defined by the effect of the treatment on outcome through each mediator. A MATE is akin to (and a special version of) an "indirect effect" in the causal mediation analysis literature (Imai et al. 2010; Acharya et al. 2016).

3. When the mediator is no longer strictly exogenous, the usefulness of the FDC de-

pends on the degree of exogeneity of the mediator. In cases where the mediator is only plausibly—but not strictly—exogenous (Conley et al., 2012), the estimate of the ATE obtained by the FDC is closer to the true value of the ATE than the estimate of the ATE obtained by a naïve regression of outcome on treatment. In cases where the mediator is deemed to be strongly endogenous, the estimate of the ATE obtained by the FDC is further from the true value of the ATE than the estimate of the ATE obtained by a naïve regression of outcome on treatment.

4. The FDC is most promising in cases where units of observations are selected into treatment on the basis of unobservables which also affect the outcome, but for which treatment intensity or non-compliance to the treatment can argued to be (as good as) randomly assigned.

Ultimately, the front-door criterion is a useful tool for applied researchers interested in causal inference with observational data. When selection into treatment is endogenous but there exists a single, plausibly exogenous mediator whereby the treatment causes the outcome, the front-door criterion can be argued to credibly identify the causal effect of treatment on outcome.

References

Acharya, A., Blackwell, M., and Sen, M. (2016) "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects," *American Political Science Review*, vol. 110, no. 3, pp. 512-529.

Angrist, J. and Kruger, A. (1995) "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, vol. 13, issue, 2, pp. 225—235.

Athey, S., Chetty, R., Imbens, G., and Kang, H. (2019) "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *NBER Working Paper No. 26463*.

Azar, O.H. (2020) "The Economics of Tipping," *Journal of Economic Perspectives*, vol. 34, no. 2, pp. 215-236.

Beaman, L., Karlan, D., Thuysbaert, B., and Udry, C. (2013) "Profitability of Fertilizer: Experimental Evidence from Female Rice Farmers in Mali," *American Economic Review*, vol. 103, no. 3, pp. 381-386.

Bellemare, M.F., and Wichman, C.J. (2020) "Elasticities and the Inverse Hyperbolic Sine Transformation," *Oxford Bulletin of Economics and Statistics*, vol. 82, no. 1, pp. 50-61.

Boyes, W.J., Mounts, W.S., and Sowell, C. (2006) "Restaurant Tipping: Free-Riding, Social Acceptance, and Gender Differences," *Journal of Applied Social Psychology*, vol. 34, no. 12, pp. 2616-2625.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020) "Are Referees and Editors in Economics Gender Neutral?," *Quarterly Journal of Economics*, vol. 135, no. 1, pp. 269-327.

Chandar, B., Gneezy, U., List, J.A., and Muir, I. (2019) "The Drivers of Social Preferences: Evidence From a Nationwide Tipping Experiment," NBER Working Paper no. 26380.

CompassRed (2019)"Want To Get a Tip As An Uber Driver? Don't Pick-Up A Shared

Ride," https://www.compassred.com/data-journal/want-to-get-a-tip-as-an-uber-driver-dont-pick-up-a-shared-ride last accessed May 26, 2020.

Conley, T.G., C.B. Hansen, and P.E. Rossi (2012) "Plausibly Exogenous," *Review of Economics and Statistics*, vol. 94, no. 1, pp. 260-272.

Davidson, R. and MacKinnon, J. G. (2004) *Econometric Theory and Methods*, Oxford University Press, New York.

Fernandez, R. and Rodrik, D. (1991) "Resistance to Reform: Status Quo Bias in the Presence of Individual-Specific Uncertainty," *American Economic Review*, vol. 81, no. 5, pp. 1146-1155.

Glynn, A.N. and Kashin, K. (2017) "Front-Door Difference-in-Differences Estimators," *American Journal of Political Science*, vol. 61, no. 4, pp. 989-1002.

Glynn, A.N. and Kashin, K. (2018) "Front-Door Versus Back-Door Adjustment With Unmeasured Confounding: Bias Formulas for Front-Door and Hybrid Adjustments With Application to a Job Training Program," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1040-1049.

Gupta, S., Z.C. Lipton, and Childers, D. (2020) "Estimating Treatment Effects with Observed Confounders and Mediators," Working Paper, Carnegie Mellon University.

Haavelmo, T. (1943) "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, vol. 11, no. 1, pp. 1–12.

Heckman, J., Pinto, R., and Savelyev, P. (2013) "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, vol. 103, no. 6, pp. 20052–2086.

Hemel, D. J. (2017) "Pooling and Unpooling in the Uber Economy," *University of Chicago Legal Forum* vol. 2017, pp. 265-286.

Imai, K., Keele, L., and Yamamoto, T. (2010) "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, vol. 25, no. 1, pp. 51-71.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011) "Unpacking the Black Box

of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review*, vol. 105, no. 4, pp. 765-789.

Imbens, G.W. (2020) "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *Journal of Economic Literature*, forthcoming.

Imbens, G.W., and Angrist, J.D. (1994) "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, vol 62, no. 2, pp. 467-475.

Kahneman, D. Knetsch, J.L., and Thaler, R.H. (1991) "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, vol. 5, no. 1, pp 193-200.

Morgan, S.L., and Winship, C. (2015) *Counterfactuals and Causal Inference*, Cambridge University Press, Cambridge, United Kingdom.

Pearl, J. (1993) "Mediating Instrumental Variables," *Technical Report R-210*.

Pearl, J. (1995) "Causal Diagrams for Empirical Research," *Biometrika*, vol. 82, no. 4, pp. 669-688.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, United Kingdom.

Pearl, J. and Mackenzie D. (2018) *The Book of Why: The New Science of Cause and Effect*, Basic Books: New York, NY.

Rosenbaum, P. and Rubin, D. (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, pp. 41-55.

Rubin, D.B. (2005) "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322-331.

Strotz, R.H. and Wold, H.O.A. (1960) "Recursive versus nonrecursive systems: An attempt at synthesis," *Econometrica*, vol. 28, pp. 417–427.

Wu, Y., Chen, X., and Jingwen, M. (2018) "Modeling Passengers' Choice in Ride-Hailing Service with Dedicated-Ride Option and Ride-Sharing Option," ICIBE' 18: Pro-

ceedings of the 4th International Conference on Industrial and Business Engineering, pp. 94-98.

Zellner A.(1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, vol. 57, issue 298, pp. 348—368.

Appendix

A1. Mediators as Instruments

In this appendix we discuss the consequences of using the exogenous mediator M, which satisfies the FDC requirements and assumptions, as an instrumental variable. Recall the usual instrumental variable (IV) setup required for the local average treatment effect (LATE) theorem to hold (Imbens and Angrist 1994): The treatment variable X is endogenous to the outcome of interest Y, but the econometrician has access to an instrumental variable Z which (i) is independent, (ii) satisfies the exclusion restriction, and is (ii) relevant. At the outset, it may seem tempting to simply use the mediator M as an instrument instead of using the front-door criterion. We show here why this is not advisable.

Assume, for illustrative purposes, that both the treatment X and mediator M are binary variables. Further assume, as is the case in our ride-sharing empirical application, that we have one-sided noncompliance. So, if $M = 1$ then $X = 1$ and if $M = 0$ then $X = 1$ or $X = 0$. In this case, there are zero never-takers and zero defiers in the sample by construction. With one-sided non-compliance, of the sort defined in this illustration, the entire sample are either compliers or always takers.

The core difference between using the mediator M as an instrument versus using the mediator M in the FDC method is that the former estimates a LATE and the latter estimates an ATE. In an ideal IV set-up, with monotonicity, the instrument removes endogenous variation driven by differences between the compliers and the never takers or always takers. The resulting LATE is the ATE for the sub-set of the sample who comply to the instrument. In an ideal FDC set-up, non-compliance is exogenous and identification fundamentally relies on this variation. Thus, the FDC calculates an ATE for the entire sample.

More specifically, although the placement of M on the path between X and Y means that M cannot be used to identify a LATE because M does not cause X, exogeneity of M in the FDC setup means that M can still be used as an IV (provided the relevance

requirement is satisfied) because M satisfies the independence requirement.[29] In practice this means that instead of causing X as in the LATE scenario, M is merely correlated with X when used as an IV. This brings a few issues to the fore. First, on the relevance front, when there is low correlation between Y and M, the latter is a weak instrument if it is used as an IV, which means that the estimate obtained therefrom collapses to that of the naïve OLS estimator. If instead M is used in an FDC setup, the regression of Y on M conditional on X yields a statistically insignificant coefficient on M, and so the estimate of the ATE goes to zero.

Second, and more importantly, when using M as an IV, the estimate of the treatment effect obtained will almost surely be different than the estimate of the (average) treatment effect obtained when using M in context of the FDC. Indeed, the usual IV setup—in which the relationship between X and the IV is given a causal interpretation—allows estimating a LATE precisely because uptake of treatment X is caused by the IV. But if one were to use M as an IV, rather than having M cause X, one would only have M be correlated with X (say, via the unobserved propensity to take up the treatment) instead of being caused by it. But then, the LATE interpretation no longer holds, since individual units are no longer induced to take up the treatment by the instrument, and it is not entirely clear what interpretation can be given to the IV estimate. Between using M in an FDC setup (and recovering the ATE of X on Y) on the one hand and using M as an IV for X (and recovering a nebulous treatment effect whose interpretation is unclear) on the other hand, the former is obviously preferable.

[29] In this case, the IV is more accurately described as a surrogate IV (Morgan and Winship 2015). See Athey et al. (2019) on the use of multiple surrogates to estimate treatment effects.

A2. Multiple Mediators–Case 2

We now show the results of a simulation that demonstrate the consequences (or lack thereof) of multiple mediators of the sort illustrated in Figure I, where multiple mediators lie on the same path from X to Y.

Our simulation setup is as follows. Let $U \sim N(0,1)$, $\epsilon_X \sim N(0,1)$, $Z_1 \sim U(0,1)$, $Z_2 \sim U(0,1)$, $\epsilon_{M1} \sim N(0,1)$, $\epsilon_{M2} \sim N(0,1)$, and $\epsilon_Y \sim N(0,1)$ for a sample size of $N = 100,000$ observations. Then, let

$$X_i = 0.5U_i + \epsilon_{Xi}, \tag{38}$$

$$M_{1i} = Z_{1i}X_i + \epsilon_{M1i}, \tag{39}$$

$$M_{2i} = Z_{2i}M_{1i} + \epsilon_{M2i}, \tag{40}$$

and

$$Y_i = 0.5M_{2i} + 0.5U_i + \epsilon_{Yi}. \tag{41}$$

As illustrated in Figure I, this fully satisfies Pearl's (1995, 2000) three criteria for the FDC to be able to estimate the average treatment effect of X on Y. By substituting equation 39 into equation 40 and substituting equation 40 into Equation 41, it should be immediately obvious to the reader that the true ATE is equal to 0.125 in our simulations.

Similar to the previous simulation analysis, we estimate several specifications. The first specification estimates the true ATE by controlling for the confounder U. The second specification estimates the ATE using the FDC approach. As the results in Table VIII show, estimates of the ATE with the FDC approach in this case are statistically invariant whether either or both M_1 and M_2 are included in the estimation procedure.

TABLE VIII: Simulation Results—Multiple Mediators, Case 2

Variables	Benchmark	Naïve	Front-Door (Both)			Front-Door (M_1 only)		Front-Door (M_2 only)	
	Y	Y	M_1	M_2	Y	M_1	Y	M_2	Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment (X)	0.127***	0.326***	0.495***	0.245***	0.201***	0.496***	0.120***	0.245***	0.202***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
Mediator (M_1)	–	–	–	–	0.003	–	0.254***	–	–
					(0.004)		(0.004)		
Mediator (M_2)	–	–	–	–	0.502***	–	–	–	0.503***
					(0.003)				(0.003)
Confounder (U)	0.501***	–	–	–	–	–	–	–	–
	(0.004)								
Intercept	-0.001	0.004	0.002	0.004	0.002	0.002	0.004	0.004	0.002
	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)
Estimated ATE	0.127***	0.326***		0.125***		0.126***		0.123***	
	(0.004)	(0.004)		(0.002)		(0.002)		(0.002)	
Observations	100,000	100,000		100,000		100,000		100,000	

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The front-door equations in columns (3) and (4) are estimated by seemingly unrelated regressions. The standard error for the front-door ATE is estimated by the delta method.

A3. Real-World Application: Experimental Replication

This section illustrates the FDC using the results of an experimental study by Beaman et al. (2013). In Table IX, we replicate results from Beaman et al. (2013), who conduct a randomized controlled trial with rice farmers in Mali. Starting from the full sample, units of observations are either assigned to a treatment group or a control group, with treated units receiving fertilizer and control units receiving no fertilizer.

In this application, we exploit as a mediator the fact that treatment intensity varies at random within the treatment group to illustrate the FDC in practice. About half of the treatment-group observations receive half of the prescribed amount of fertilizer, the remainder of the treatment-group observations receiving the full prescribed amount of fertilizer, and the control-group observations receiving none of the prescribed amount of fertilizer.

As one would expect from the derivations in Section 2, the results in Table IX show that the ATEs obtained by the FDC are all statistically indistinguishable from the benchmark ATEs. For example, considering the average rate of fertilizer use among the control group is 0.32, the benchmark estimate (in column 1) suggests that receiving free fertilizer increases the use of fertilizer over twofold. The FDC approach roughly replicates (in column 2) this benchmark estimate. The similarity between the benchmark and FDC estimates persist for the the quantity of fertilizer use (columns 3 and 4) and fertilizer expenses (columns 5 and 6). The results show that receipt of free fertilizer leads to increases in the use of fertilizer at both the extensive and intensive margins and reduces fertilizer expenses.

A few remarks are in order. First, the real-world results in this section are most useful for highlighting the potential of the FDC approach in estimating treatment effects in settings where each of the conditions hold. Of course, since Beaman et al. (2013) assign treatment experimentally, the FDC approach is not necessary to estimate treatment effects in that context.

TABLE IX: Empirical Illustration — Rice Production and Fertilizer Use in Mali

	Use of Fertilizer		Fertilizer Quantity		Fertilizer Expenses	
	(1)	(2)	(3)	(4)	(5)	(6)
Benchmark	0.639***		27.24***		-2,717.1***	
	(0.033)		(3.568)		(464.6)	
Front-Door		0.603***		26.64***		-2,605.3***
		(0.030)		(3.002)		(389.7)
Observations	378	378	378	378	373	373

Notes: Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All columns include the same control variables as in Beaman et al. (2013). Columns (1), (3), and (5) represent benchmark OLS estimates of the ATEs of receiving fertilizer on the outcome variables. The estimates in columns (1), (3), and (5) replicate the findings of Beaman et al. (2013) except that the original study differentiates between two treatment groups defined by intensity of treatment. Columns (2), (4), and (6) represent seemingly unrelated regression estimates of the front-door criterion ATEs. Standard errors in columns (2), (4), and (6) are estimated by the delta method.

Second, in that real-world experimental case, there is no need to condition on the treatment variable (i.e., X) when estimating the effect of the mediator (i.e., M) on the outcome (i.e., Y) since the random assignment of treatment already removes any back-door path between Y and M. In fact, needlessly conditioning on the treatment variable in an experimental setting leads to bias in the front-door estimate, due to violating the data requirement that $P(X_i|M_i) > 0$.

A4. Treatment Totally Defined by the Mediator—Exogenous Treatment

In section 4.3 we discussed Pearl's additional condition, or data requirement, for the FDC method: that $P(X_i|M_i) > 0$. Through the use of simulated data, we demonstrated that violating this assumption is really only problematic when the treatment is exogenous. In this section, we show results using the experimental data of Beaman et al. (2013) to further highlight this detail.

The results in Table X revisit the data in Table IX. Here, however, odd-numbered columns report the correct ATEs, and even-numbered columns report biased ATEs. These results show that when treatment is exogenous, it is not necessary to include treatment into the second-stage FDC regression because there are no back-door paths from U to X. In fact, conditioning on treatment could lead to biased estimates when using the FDC method.

Why does the treatment need to be omitted from Equation 8 when the assumption that $P(X_i|M_i) > 0$ is violated and there are no unobserved confounders? In such cases, the variation in X is already accounted for in the variation in M. Indeed, when $M_i > 0$, we know $X_i = 1$, and when $M_i = 0$, we know $X_i = 0$. Although it is certainly possible to estimate Equation 8 when the treatment is totally defined by the mediator, both with and without unobserved confounders, it is only in the former case that the inclusion of both M and X as regressors on the left-hand side of Equation 8 will return an unbiased ATE.

TABLE X: Over-Controlling for the Treatment — Fertilizer Use in Mali

	Use of Fertilizer		Fertilizer Quantity		Fertilizer Expenses	
	(1)	(2)	(3)	(4)	(5)	(6)
Benchmark Front-Door ATE	0.603***		26.64***		-2,605.3***	
	(0.030)		(3.002)		(389.7)	
Over-Controlled Front-Door ATE		0.009		17.580***		-882.72
		(0.055)		(5.920)		(776.90)
Observations	378	378	378	378	377	377

Notes: Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All columns include the same control variables as in Beaman et al. (2013). Columns (1), (3), and (5) represent benchmark seemingly unrelated regression FDC estimates of the ATEs of receiving fertilizer on the outcome variables. Columns (2), (4), and (6) represent seemingly unrelated regression estimates of the front-door criterion ATEs which over-control for treatment in the outcome regression of the FDC setup. Standard errors are estimated by the delta method.